Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

- Count is higher when it is not a holiday
- Count is higher in May, July, August, September, June in that order. We have good demand starting from april till September. Low demand on October, November, December, January, February.
- Summer and Fall has higher demand than Spring and Winter. Spring has the lowest demand.
- clear whether has higher demand than misty weather and rainy weather
- Demand is the least on sundays.
- Demand is more on working days
- 2019 has higher demand than 2018.

2. Why is it important to use drop_first=True during dummy variable creation?

We will encounter a problem called Dummy variable trap. If we have n dummy variables, we need to use n-1 dummy variables. If we use n dummy variables, we will have multicollinearity problem. one dummy variable can be predicted(perfectly correlated) using all other dummy variables. So we need to drop one dummy variable.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

- apparent_temparature is highly correlated with count variable.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?

- Linearity assumption with partial residual plots
- Normality for residuals with Q-Q plots
- Homoscedasticity with Residuals vs Fitted plot
- No autocorrelation with Durbin-Watson test

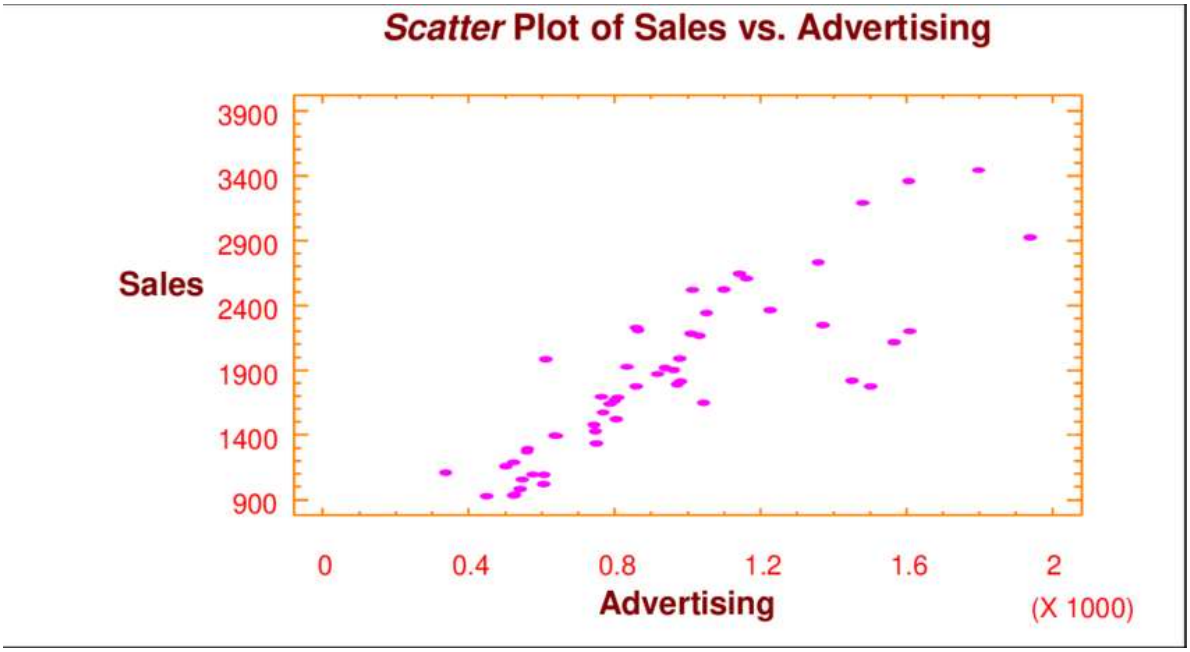
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

- 2019 0.230502
- Light Snow/Rain -0.248777
- temp 0.509836
- year, weather_code and temperature are the top 3 features contributing significantly towards explaining the demand of the shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

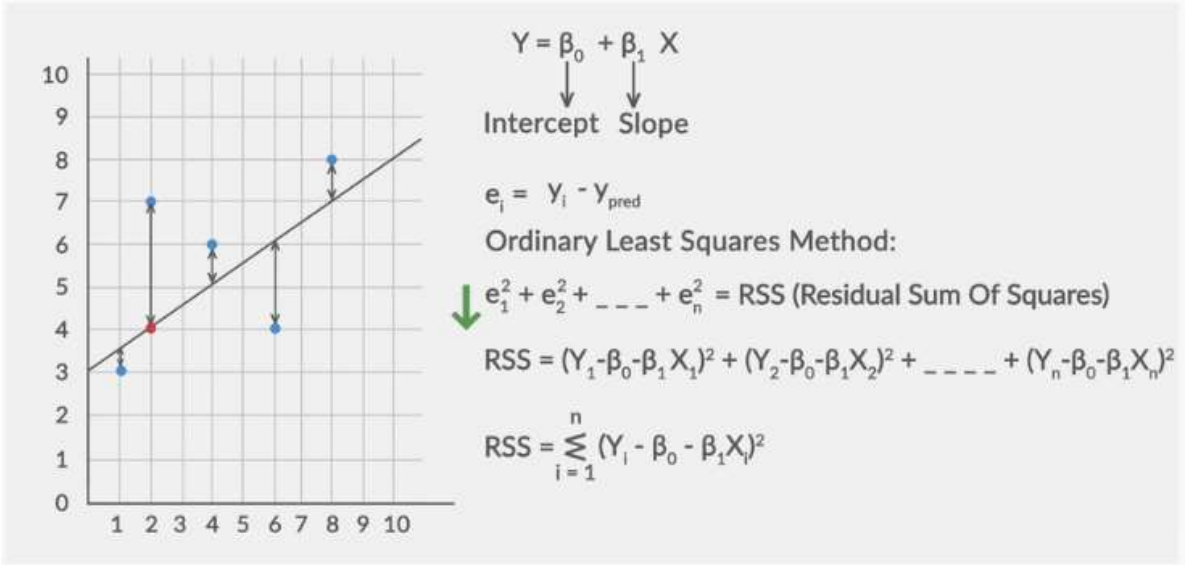
Linear regression is a supervised learning method for predicting continous variables. Examples are price of a house, score of a student, demand of a product etc. Since this is a supervised learning algorithm we need previous data for training a model. Linear regression is a predictive modeling technique which explains the difference between the actual and predicted values of the dependent variable(target variable) using the independent variables(predictior variables).



1. Simple linear regression: explains the relationship between dependent variable and the independent variable using a straight line.

$$y = \beta_0 + \beta_1 x_1$$

The best-fit line is found by minimising the expression of RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot. Residuals for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable: #



The strength of the linear regression model can be assessed using 2 metrics:

1. R² or Coefficient of Determination
2. Residual Standard Error (RSE) R² or Coefficient of Determination You also learnt an alternative way of checking the accuracy of your model, which is R2 statistics. R2 is a number which explains what portion of the given data variation is explained by the developed model. It always takes a value between 0 & 1. In general term, it provides a measure of how well actual outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model, i.e. expected outcomes. Overall, the higher the R-squared, the better the model fits your data. Mathematically, it is represented as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$R^2 = 1 - \frac{RSS}{TSS}$$

Fig 5 - R-squared RSS (Residual Sum of Squares): In statistics, it is defined as the total sum of error across the whole sample. It is the measure of the difference between the expected and the actual output. A small RSS indicates a tight fit of the model to the data. It is also defined as follows:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

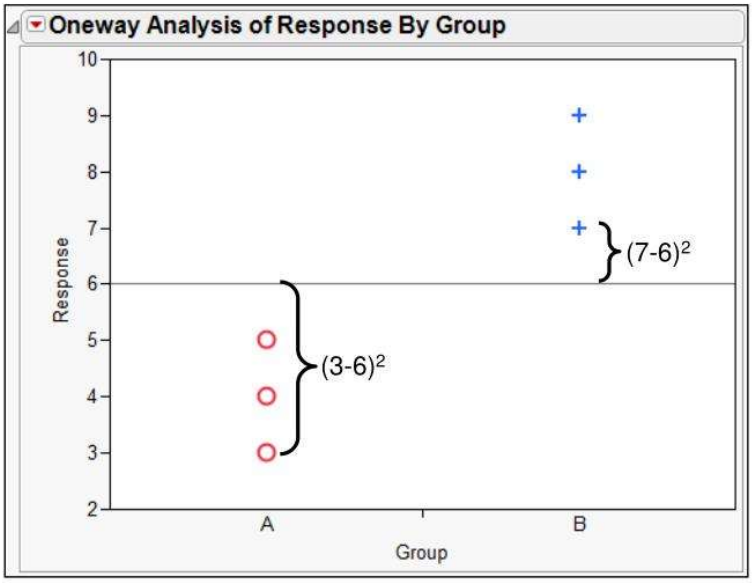
TSS (Total sum of squares): It is the sum of errors of the data points from mean of response variable. Mathematically, TSS is:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Importance of RSS/TSS: Think about it for a second. If you know nothing about linear regression and still have to draw a line to represent those points, the least you can do is have a line pass through the mean of all the points as shown below. #

jmp

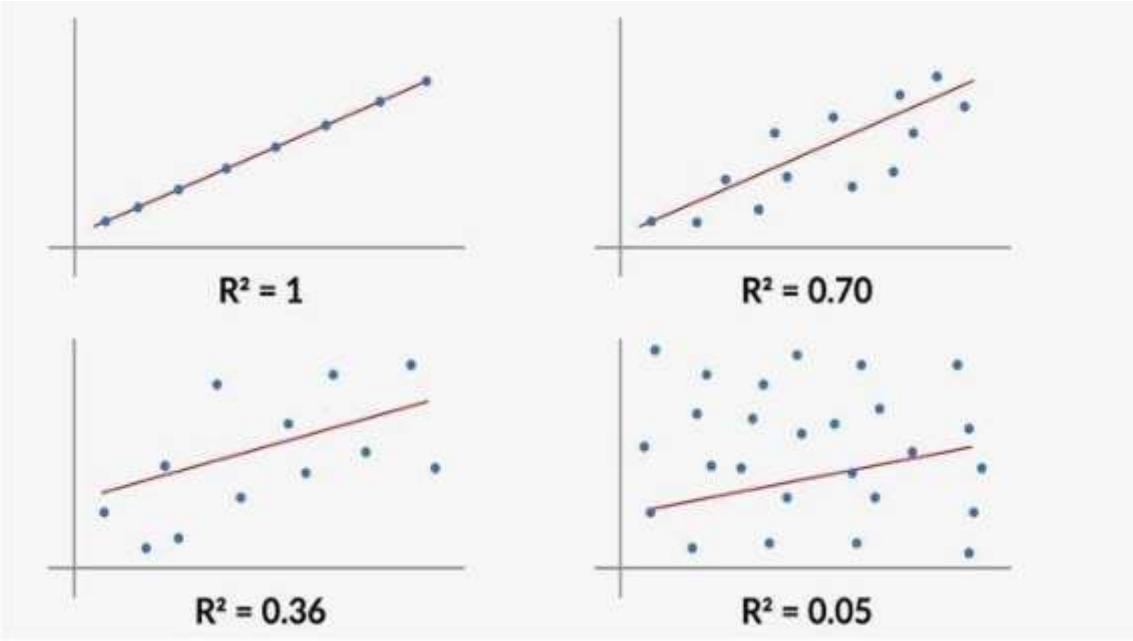
Total Sum of Squares



$$SS_T = (3-6)^2 + (4-6)^2 + (5-6)^2 + (7-6)^2 + (8-6)^2 + (9-6)^2 = 28$$

5

In Graph 1: All the points lie on the line and the R2 value is a perfect 1 In Graph 2: Some points deviate from the line and the error is represented by the lower R2 value of 0.70 In Graph 3: The deviation further increases and the R2 value further goes down to 0.36 In Graph 4: The deviation is further higher with a very low R2 value of 0.05 #



Linear regression makes the following assumptions: # 1. Linear relationship between dependent and independent variables # 2. Normality of residuals(error terms are normally distributed) # 3. Homoscedasticity of residuals(error terms have constant variance) # 4. No autocorrelation of residuals(error terms are independent of each other) There should be not

visible patterns in error terms. In linear regression, hypothesis testing is often conducted to assess the significance of the beta coefficients (the slopes) associated with each predictor variable.

Steps for Hypothesis Testing of Beta Coefficients:

1. **Null and Alternative Hypotheses:**
   - **Null Hypothesis ((H_0)):** (H_0) states that there is no relationship between the predictor variable and the target variable. It implies the beta coefficient is zero.
   - **Alternative Hypothesis ((H_1) or (H_a)):** (H_1) contradicts the null hypothesis, suggesting that there is a relationship between the predictor and the target, meaning the beta coefficient is not zero.
2. **Test Statistic:**
   - The test statistic used for the beta coefficient is often the t-statistic. It measures the number of standard deviations the coefficient estimate is from zero. The formula for the t-statistic is:

     $$t = \frac{\text{coefficient estimate}}{\text{standard error of the coefficient}}$$

     or

     $$t = \frac{\beta_1}{SE(\beta_1)}$$
3. **Degrees of Freedom and Critical Value:**
   - The degrees of freedom for the t-distribution in this case are usually (n - p - 1) (where (n) is the number of observations and (p) is the number of predictors including the intercept).
   - The critical value for the t-test is obtained from the t-distribution table or calculated using statistical software, considering the chosen significance level (e.g., 0.05, 0.01).
4. **Decision Rule:**
   - If the calculated t-statistic is greater than the critical value (for a given significance level), then we reject the null hypothesis. This implies that the beta coefficient is significant.
   - If the calculated t-statistic is less than the critical value, we fail to reject the null hypothesis, indicating that the beta coefficient is not statistically significant.
5. **p-value:**
   - The p-value associated with the t-statistic indicates the probability of observing the coefficient estimate if the null hypothesis were true. A lower p-value (usually below the chosen significance level) suggests stronger evidence against the null hypothesis.

Interpretation:

- **Significant Coefficient:** If the t-statistic is large and the associated p-value is less than the chosen significance level (e.g., 0.05), it indicates that the predictor variable has a statistically significant effect on the target variable.

- **Non-significant Coefficient:** If the t-statistic is small and the p-value is greater than the chosen significance level, there is insufficient evidence to reject the null hypothesis, suggesting that the predictor variable may not have a significant effect on the target variable.

This hypothesis testing helps in determining which predictors are contributing significantly to the model and which ones might be less influential.

After determining that the coefficient is significant, using p-values, you need some other metrics to determine whether the overall model fit is significant. To do that, you need to look at a parameter called the F-statistic. So, the parameters to assess a model are:

1. t statistic: Used to determine the p-value and hence, helps in determining whether the coefficient is significant or not
2. F statistic: Used to assess whether the overall model fit is significant or not. Generally, the higher the value of F statistic, the more significant a model turns out to be
3. R-squared: After it has been concluded that the model fit is significant, the R-squared value tells the extent of the fit, i.e. how well the straight line describes the variance in the data. Its value ranges from 0 to 1, with the value 1 being the best fit and the value 0 showcasing the worst.

Multiple linear regression is a statistical technique to understand the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X. Consider our previous example of sales prediction using TV marketing budget. In real life scenario, the marketing head would want to look into the dependency of sales on the budget allocated to different marketing sources. Here, we have considered three different marketing sources, i.e. TV marketing, radio marketing, and newspaper marketing. You need to consider multiple variables as just one variable alone might not be good enough to explain the feature variable, in this case, Sales. The table below shows how adding a variable helped increase the R-squared that we had obtained by using just the TV variable. So we see that adding more variables increases the R-squared and it might be a good idea to use multiple variables to explain a feature variable. Basically:

1. Adding variables helped add information about the variance in Y!
2. In general, we expect explanatory power to increase with increase in variables Hence, this brings us to multiple linear regression which is just an extension to simple linear regression. The formulation for multiple linear regression is also similar to simple linear linear regression with the small change that instead of having beta for just one variable, you will now have betas for all the variables used. The formula now can be simply given as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon$$

Apart from the formula, a lot of other ideas in multiple linear regression are also similar to simple linear regression, such as:

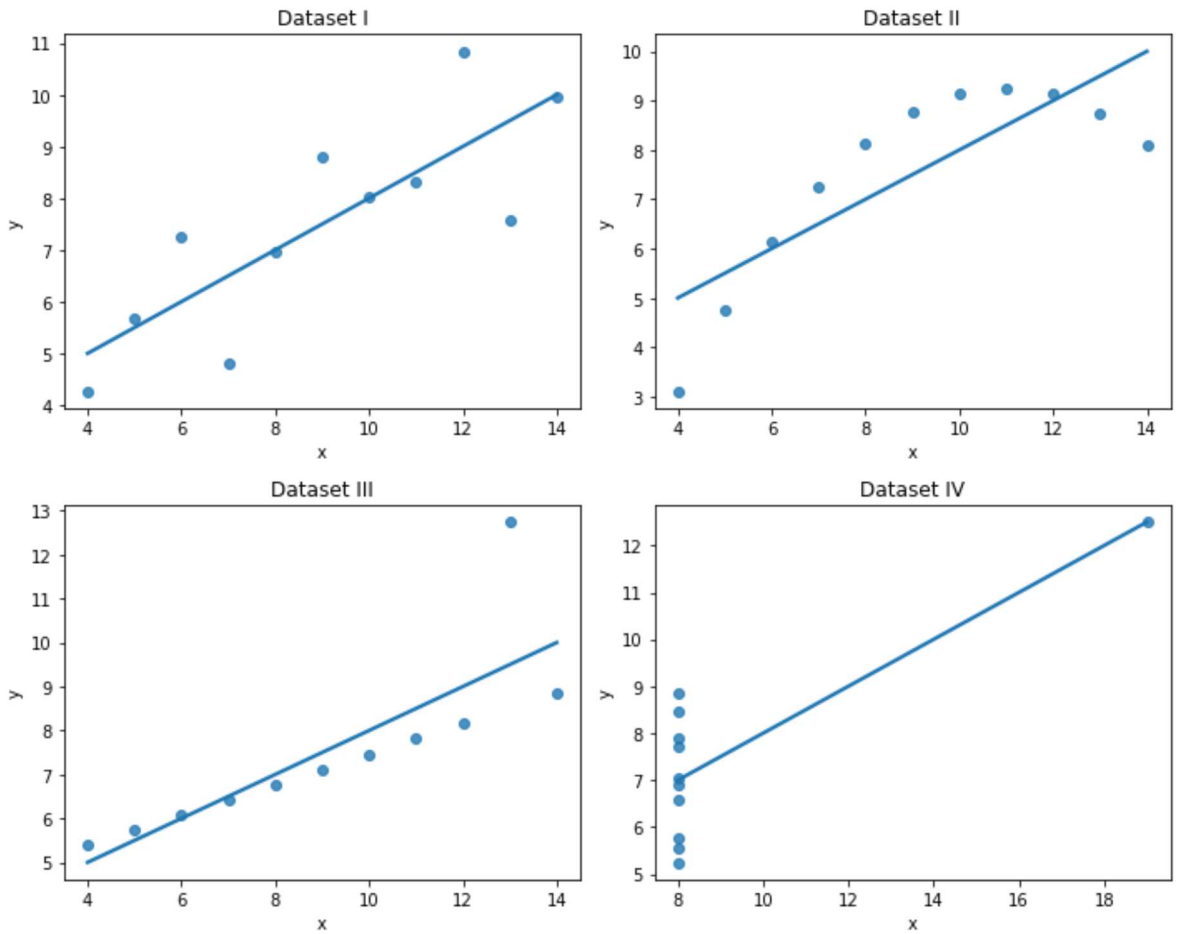
2. Explain the Anscombe’s quartet in detail.

In 1973, the statistician Francis Anscombe used a clever set of bivariate datasets (now known as Anscombe’s quartet) to illustrate the importance of graphing data as a component of statistical analyses. In his example, each of the four datasets yielded identical regression coefficients and model fits, and yet when visualized revealed strikingly different patterns of covariation between x and y.

Upon visual inspection, it becomes immediately clear that these datasets, while seemingly identical according to common summary statistics, are each unique. This is the power of effective data visualization: it allows us to bypass cognition by communicating directly with our perceptual system.

In [3]:

```
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd

# Anscombe's Quartet Data
data = sns.load_dataset('anscombe')

# Separate datasets I, II, III, IV
dataset_I = data[data['dataset'] == 'I']
dataset_II = data[data['dataset'] == 'II']
dataset_III = data[data['dataset'] == 'III']
dataset_IV = data[data['dataset'] == 'IV']

# Plotting
fig, axes = plt.subplots(2, 2, figsize=(10, 8))

for i, dataset, label in zip(range(4), [dataset_I, dataset_II, dataset_II
    ax = axes[i // 2, i % 2]
    sns.regplot(x='x', y='y', data=dataset, ax=ax, ci=None)
    ax.set_title(f'Dataset {label}')

plt.tight_layout()
plt.show()
```



3. What is Pearson's R?

Karl Pearson developed the statistical measure now known as Pearson's r, or the Pearson product-moment correlation coefficient, around the turn of the 20th century. Briefly stated, Pearson's r measures the covariance of two variables in terms of their standard deviations. In other words, leaving aside the units and scale of the two variables, is growth in one variable consistently reflected in comparable increase or decrease in the other. It is given by

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where x and y are the two variables,
n is the number of observations,
x_i and y_i are the individual observations, and
$\bar{x}$ and $\bar{y}$ are the means of x and y, respectively.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling

and standardized scaling?

Feature scaling is an important aspect to consider when dealing with a lot of independent variables in a model. When these variables are on very different scales, it can lead to a model with coefficients that are difficult to interpret. There are two main reasons why we need to scale features: ease of interpretation and faster convergence for gradient descent methods.

Two popular methods for scaling features are standardizing and MinMax scaling. In standardizing, the variables are scaled in such a way that their mean is zero and standard deviation is one. In MinMax scaling, the variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

It is important to note that scaling only affects the coefficients and none of the other parameters like t-statistic, F statistic, p-values, R-square, etc.

Standard scaling:

$$z = \frac{x - \mu}{\sigma}$$

where x is the original feature value, μ is the mean of the feature values, σ is the standard deviation of the feature values, and z is the scaled feature value.

Min-max scaling:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

where x is the original feature value, x_{min} is the minimum feature value, x_{max} is the maximum feature value, and x' is the scaled feature value.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

When Variance inflation factor(VIF) is infinite it indicates perfect multicollinearity between independent variables. Perfect multicollinearity means One variable can be predicted by other variables. We need to drop the vairables that cause multicollinearity

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A Q-Q plot (quantile-quantile plot) is a probability plot, a graphical method for comparing two probability distributions by plotting their quantiles against each other. A Q-Q plot is a plot of the

quantiles of two distributions against each other, or a plot based on estimates of the quantiles. The pattern of points in the plot is used to compare the two distributions. The main step in constructing a Q-Q plot is calculating or estimating the quantiles to be plotted.

In a Q-Q plot, if the two distributions being compared are similar, the points in the plot will approximately lie on the identity line y = x. If the distributions are linearly related, the points in the Q-Q plot will approximately lie on a line, but not necessarily on the line y = x. Q-Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions. A Q-Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions. Q-Q plots can be used to compare collections of data or theoretical distributions. The use of Q-Q plots to compare two samples of data can be viewed as a non-parametric approach to comparing their underlying distributions.

In linear regression, Q-Q plots are used to check the normality assumption of the residuals. If the residuals are normally distributed, the Q-Q plot will be approximately a straight line. If the residuals are not normally distributed, the Q-Q plot will deviate from a straight line. This indicates that the normality assumption of the residuals has been violated, which can affect the validity of the linear regression model.